The Sapphire Approach to Mining Science Data

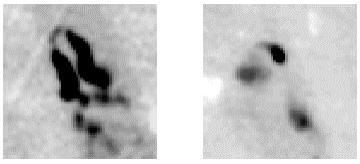
Chandrika Kamath

Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
May 8, 2002

Presentation at the TriValley Software Showcase

MIT's Technology Review (Jan'01) - data mining is a 'top ten' emerging technology

- **Data mining: The semi-automatic discovery of patterns, associations, anomalies, and statistically significant structures in data**
- **Pattern recognition: The discovery and characterization of patterns**
- **Pattern: An ordering with an underlying structure**
- **Feature: Extractable measurement or attribute**



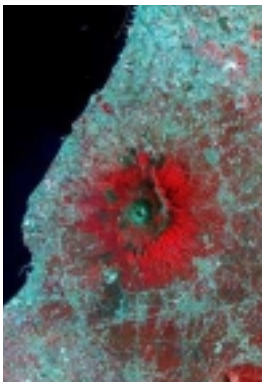
FIRST galaxy images

Pattern: Galaxies with a bent-double morphology
Features: Number of "blobs"
Spatial relationship between blobs (distances and angles)

CK 2

Data mining: multi-disciplinary field with a broad applicability

- **Has several applications**
  - market basket analysis
  - customer relationship management
  - fraud detection
  - network intrusion detection
  - non-destructive evaluation
  - astronomy (look up data)
  - remote sensing (look down data)
  - text and multi-media mining
  - medical imaging
  - automated target recognition, biometrics, …
- **Combines ideas from several different fields**



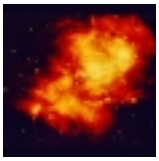
➔ **Data mining brings together the mature offshoots of technologies at a time when we are ready to exploit them.**

CK 3

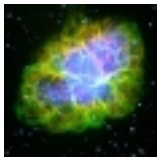
We need better data analysis to realize the full potential of enhanced data collection

- **Data obtained from experiments, observations, and simulations**
- **Science data: massive and complex**
  - **multi-sensor, multi-spectral, multi-resolution**
  - **spatio-temporal**
  - **high-dimensional**
  - **images, text, structured and unstructured meshes**
  - **contaminated with noise**

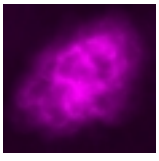
X-ray



Infrared



Optical



Radio

➔ **Visual data analysis for massive data sets is impractical given its subjective nature and human limitations in absorbing detail**

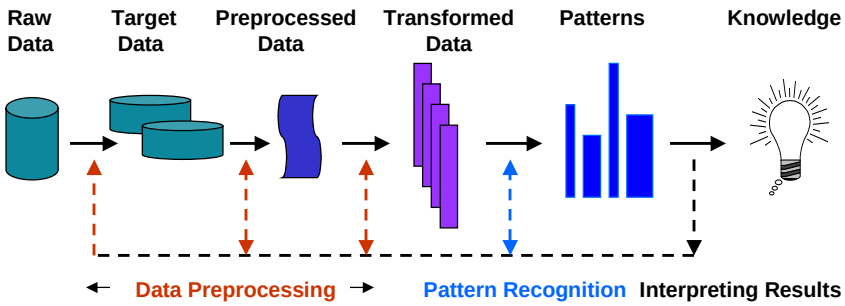
CK 4

Sapphire: a research project in scientific data mining

- Started in Oct. 98 with a three-fold focus
  - research in robust, accurate, scalable algorithms
  - incorporate the research into parallel, portable, software modules within a flexible system architecture
  - application of the software to practical problems
- A multi-disciplinary team
  - researchers and software developers
- Details, including publications, at
  - http://www.llnl.gov/casc/sapphire

CK 5

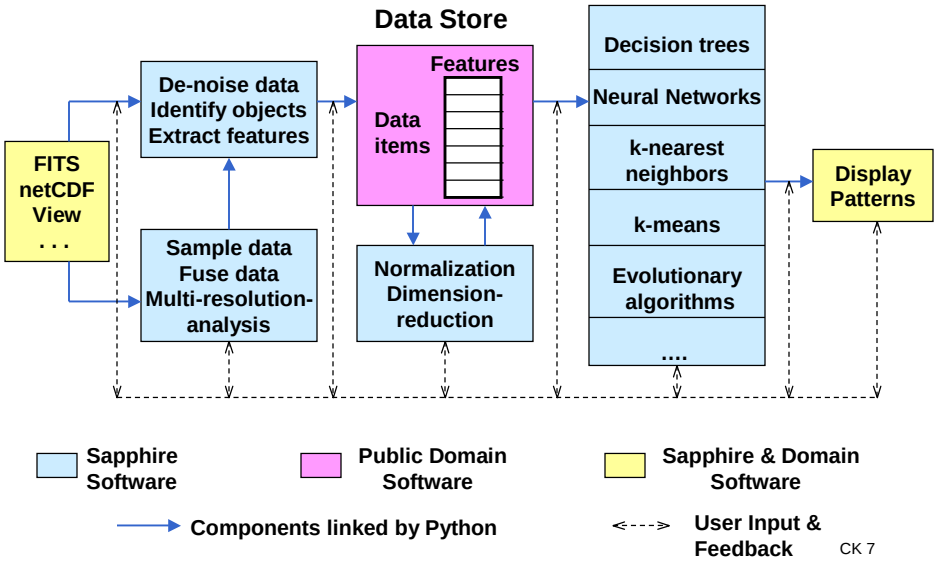
The Sapphire view of data mining - from a Terabyte to a Megabyte



An iterative and interactive process

CK 6

The Sapphire system architecture: flexible, portable, scalable



CK 7

Our research focus is on the compute-intensive parts of data mining

- Techniques for de-noising data
  - wavelet-based statistical techniques
- Improve performance of decision tree classifiers
  - faster and more accurate
  - use evolutionary algorithms and ensembles
- Current research topics
  - dimension reduction
  - robust image processing
  - scalable neural networks and clustering algorithms

CK 8

Sapphire software: Version 1.0.0 released in September 2001 (C++, serial version)

Ensembles of Classifiers
Bagging, AdaBoost, Arcx4, ASPEN

Scientific Data Processing
Several Wavelets in 1,2,3-D
single/double precision
14 linear, non-linear filters
Wavelet denoising options:
6 shrinkage rules
4 shrinkage functions

Decision Trees
7 Split Criteria
Split Finders
Pruning
Creation of decision tree
Application of decision tree

Simple Classifiers
Naïve Bayes
Gibbs

Toolbox
Sampling
Random Number Generators
Smart Pointers
1,2,3-D Vectors
1,2,3-D Boxes

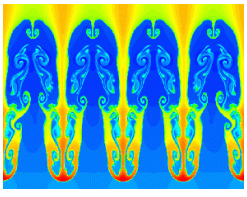
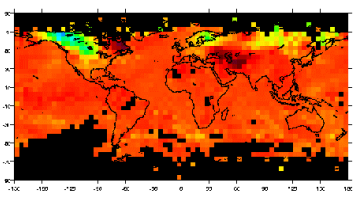
Domain Information
FITS Data
VIEW Data
Regular Data
Feature Vectors

Evolutionary Algorithms
Many options for selection, crossover, mutation, initialization, and replacement

CK 9

We are analyzing data from simulations and observations

- Classifying galaxies with a bent-double morphology
- Identifying coherent structures in turbulent flow
- Separating El Niño and volcano signals from the earth's observed temperature
- Understanding land-use in remotely-sensed data

CK 10

Where are we going from here?

- More robust, accurate, scalable algorithms
  - for pre-processing and pattern recognition
- Newer data types
  - video and multi-media
  - multi-sensor data
- More complex problems
  - dynamic tracking in video
  - mining text, audio, video, images

CK 11

Summary and contact information for more details

- Sapphire: mining massive complex data sets
  - research
  - software
  - application
- Main focus is science data, though applicable to other data sets as well
- Patents filed on software and algorithms
- More information at http://www.llnl.gov/casc/sapphire
- Contact me at kamath2@llnl.gov

UCRL-PRES-148257: This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract no. W-7405-Eng-48.

CK 12